Selection bias and auditing policies on insurance claims

Jean Pinquet

Université de Paris X, U.F.R. de sciences économiques,
200 avenue de la République 92001 NANTERRE CEDEX
FRANCE
Tél: (33)1 40 97 47 58 Fax: (33)1 40 97 59 73
E-Mails: pinquet@u-paris10.fr ; jpinquet@free.fr

with Mercedes Ayuso & Montserrat Guillén

Abstract

Statistical models which lead to optimal auditing policies are subject to selection bias. Indeed, there is a severe discrepancy between the range of derivation of the risk model (the audited claims) and its range of application (the incoming fraud suspicious claims). This paper applies a classical statistical model (bivariate probit model with censoring) which allows to assess and to counteract selection bias. The model is applied on a Catalan data base, and consequences are drawn on the efficiency of audit policies.

Keywords

Bivariate probit model with censoring. Bivariate Gaussian copulas.

1

1 Introduction

Auditing policies derived from statistical analysis and applied to insurance claims face a major selection bias problem. A score assessing fraud risk is derived from a regression analysis on audited claims, and is usually applied to all the incoming claims in order to select those which are then recommended for audit. This discrepancy between the range of derivation of the risk model (i.e. the audited claims) and its range of application (the incoming claims) creates the selection bias.

Random auditing of claims is the basic strategy which makes it possible to counteract selection bias. A pure random auditing strategy consists in selecting claims at random, then in auditing these claims. Thanks to this controlled experiment, all the fraudulent claims in the selected sample will be identified as such. The estimation of a single fraud equation in this sample provides an estimated fraud probability for incoming claims which is not subject to selection bias.

Random auditing is partly carried out on the data base which we investigated. Twenty percent of the claims are thus selected and recommended for audit. However, only one claim out of five is eventually audited in this population (see Section 2 for more details about auditing processes). Selection bias is avoided only if all the non-audited claims can be considered to be definitely non-fraudulent. Optimal auditing strategies are then designed from the estimation of a fraud equation on the audited claims (see Ayuso, Guillén et al. (2004) for derivations with the data base used in this paper).

Random auditing is far from being used by all insurance companies. Without such a policy, the effect of selection bias on the incoming claims can easily be anticipated. The experts take an audit decision from claim characteristics recorded in the company files, but they are also able to capture idiosyncrasies in fraud distributions which are not summarized by the observable information. Given observable characteristics, fraud risk is expected to be less important for a claim exempted from audit by the expert than for another one checked for fraud. A fraud risk model derived without taking this selection problem into account would then overestimate fraud probabilities for the incoming claims.

Statistical models can offset this bias by using a two equation model. An audit equation is estimated on all the claims together with another one related to fraud, which is estimated only on the audited claims. From a joint distribution of the random components in each equation, we can condition fraud risk not only on claim characteristics but also on whether or not the claim was

2

retained for audit by the claims adjusters. To our knowledge, selection models have so far not been applied to claim auditing.

Selection models can be designed with or without censoring for the variable of interest. In our context, data are censored since fraud is checked on the audited claims only. Selection models with censoring (i.e. those for which the variable of interest is observed only on the selected individuals) are widely used in statistical and econometric literature. As regards the pair: selection variable and variable of interest, we can cite

- Heckman (1979) for the seminal paper on this topic, and an application to the pair: participation in a labor market - hourly wage. The variable of interest is defined from a linear model.

- Hausman, Wise (1979) for a panel data model with endogenous attrition. The selection variable refers to attrition and the variable of interest is a wage level.

- Credit scoring models (Hand and Henley (1997)). The problem discussed is very similar to the one addressed in this paper, since the variable of interest is binary. The dependent variable refers to credit default, and selection means acceptance of the loan demand. Statistical analysis of selection bias in this context is often referred to as rejection inference.

Selection does not always imply censoring for the variable of interest. In dealing with public policy evaluation, selection means participation in a program, an educational program for instance. In this case a variable of interest such as the wage level is observed for all the individuals. Statistical analysis leads to a "difference in differences" approach (see Rosenbaum, Rubin (1983), Heckman (2000)).

Closer to the subject dealt with in this paper is the two equation model proposed by Chiappori and Salanié (2000) which tests for asymmetric information in insurance markets. Two binary variables are explained on insurance contracts, the coverage level and an accident indicator. Data are analyzed with a bivariate probit model, and the nullity of the correlation coefficient is tested for. This approach is used on our two equation model on audit and fraud. The estimated correlation coefficient is expected to be positive since it reflects the ability of claims adjusters to assess hidden characteristics in fraud distributions.

The paper is organized as follows. Section 2 explains how fraud detection strategies for automobile insurance claims are implemented in the insurance companies. Section 3 describes the data base. Section 4 presents the bivariate probit model and its application to the correction of selection bias. We consider a natural extension of the single equation probit model on the fraud variable (see Belhadji, Dionne, Tarkani (2000) and Artis, Ayuso, Guillén (2003) for the inclusion of misclassification risk in the fraud equation). Our claims data base is split into two populations. Claims selected at random (one out of five) are recommended for audit, whereas there is no specific recommendation for the other ones. We will use the latter population as the working sample since these claims are subject to selection bias. The claims selected at random will be used as a holdout sample in order to assess the efficiency of the statistical model. The estimated correlation coefficient in the bivariate probit model is found to be positive on our data. This means that, if we control for observable information on claims, fraud risk is lower for claims exempted from audit by the experts.

Optimal auditing policies which take into account selection bias are presented in Section 5, and concluding remarks appear in Section 6.

2 Fraud detection strategies on automobile insurance claims and selection bias issues

Applying auditing strategies to fraud detection in the insurance market has not been a general practice, but insurers are beginning to realize the advantages of controlling this risk. In most European countries, when an insurer finds a claim which shows evidence of fraud, an agreement with the policyholder is usually reached. Most settlements end up not renewing the policy and others lead to a reduction of the claim compensation. This situation is in contrast with the highly litigious US system. However, even if very few fraudulent claims are discussed in court, many European insurers are beginning to implement strategies to control fraud.

Let us formalize now the selection bias issue. Suppose first that all the incoming claims can be considered for audit. The selection bias issue can be formalized in the following way: A and F denote the binary variables related to audit and fraud, and x is the vector of variables which describe the claim. A statistical model assessing fraud risk is derived on the audited claims and

4

estimates probabilities of the type $P(F = 1 \,|\, A = 1,\, x) = E(F \,|\, A = 1,\, x)$. Now an audit policy induced by this model is applied on the incoming claims, and uses the probabilities $P(F = 1 \,|\, x) = E(F \,|\, x)$. Selection bias is a consequence of the confusion between conditional and unconditional probabilities.

In reality all the incoming claims are not likely to be audited. A description of the audit process will explain why. All the incoming claims are in the first place checked by the adjusters. They consider the causes of car damages, circumstances of the accident and policy characteristics. A claim the adjuster does not find suspicious will be exempted from audit and settled routinely whatever the audit strategy. In what follows, audit means that the claim is transferred to a Special Investigation Unit, referred to as an SIU. This unit provides a further assessment of the claim and decides whether to consider it as fraudulent or not.

Hence a variable related to fraud suspicion is created by the adjuster before the audit decision, and we denote it as S. For this binary variable, we have

$$S = 0 \Rightarrow A = 0. \tag{1}$$

The condition given in (1) simply means that the audit decision is nested inside the selection decision induced by S. Since both variables S and A are binary, this condition amounts to $A \leq S$. A claim for which $S = 1$ will be referred to later as suspicious of fraud. Conversely, $S = 0$ if the first screening reveals no fraud suspicion. Another reason (although less frequent) which implies $S = 0$ is that no loss is expected, for instance if the insurer of the third party pays for the car damages or if the deductible exceeds the claim cost.

Since only the claims suspected of fraud are likely to be audited, selection bias now reflects the confusion between $P(F = 1 \,|\, A = 1,\, x)$ and $P(F = 1 \,|\, S = 1,\, x)$, under the condition given in (1). This bias increases with the proportion of suspicious claims which are not audited, and is eliminated if $A = S$ in the sample. Such a sample is created by a random auditing strategy, designed in the following way. First, a population of claims is selected at random, and the treatment (using the vocabulary of experience plans) is to audit all the suspicious claims. This pair of actions (selection of claims at random, plus treatment) provides a controlled experiment which eliminates selection bias if the fraud equation is estimated from these claims.

The audit decision for suspicious claims is transferred to the adjusters if there is no random auditing strategy. An expert takes this decision from an implicit trade-off between the estimated audit cost and gain from fraud

detection. Most of the suspicious claims are not audited if the audit decision is left to the experts (see for instance Section 3).

To sum up, Figure 1 describes the links between the three binary variables S, A and F, depending on the audit strategy.
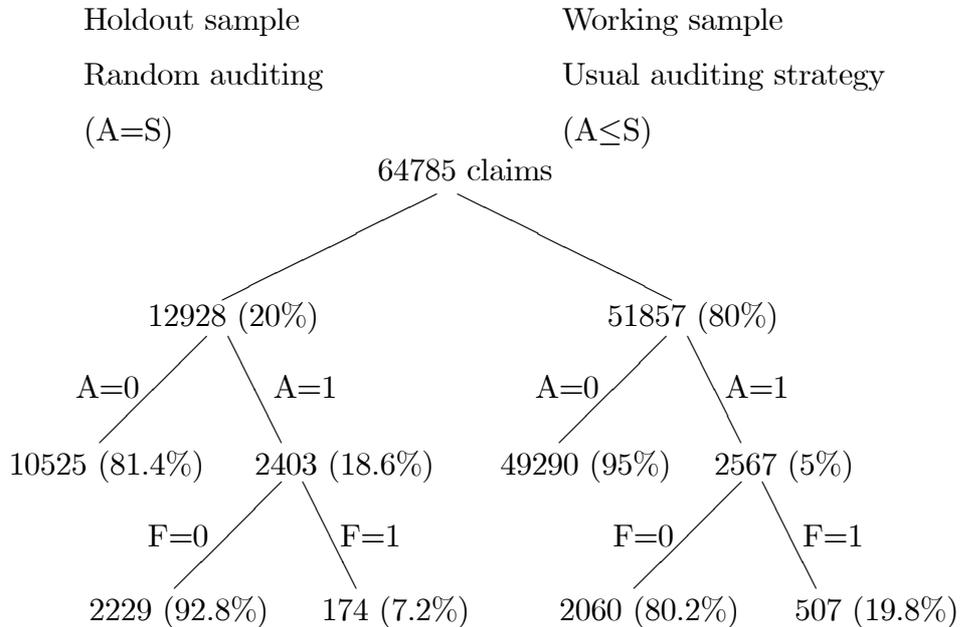
Figure 1: The binary variables S, A and F and audit strategies

Random auditing strategy $(A = S)$

Usual auditing strategy $(A \leq S)$
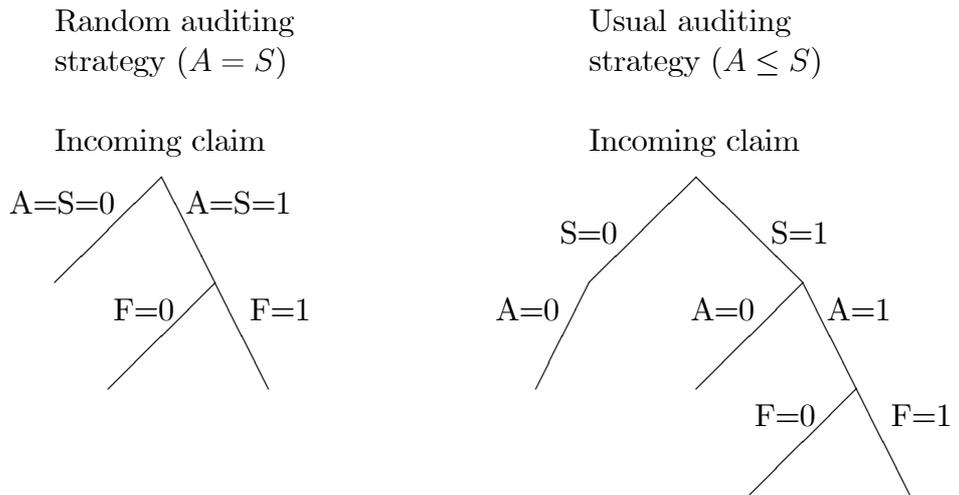
3 Presentation of the data base

The claims data base belongs to an insurance company operating in Spain. The claims are linked to motor insurance contracts and were reported during the year 2000. The structure of the data base is described in Figure 2. For each node of the tree are given the number of claims and the weight with respect to the parent node.

Figure 2: The claims data base

Holdout sample   Working sample

Random auditing   Usual auditing strategy

(A=S)   (A$\leq$S)

64785 claims

12928 (20%)   51857 (80%)

A=0   A=1   A=0   A=1

10525 (81.4%)   2403 (18.6%)   49290 (95%)   2567 (5%)

F=0   F=1   F=0   F=1

2229 (92.8%)   174 (7.2%)   2060 (80.2%)   507 (19.8%)

Let us give comments on these numbers of claims, depending on their type. The proportion of claims selected for random auditing is 20%. If random auditing is performed thoroughly, all the claims suspected of fraud are audited, which means that their weight in the sample is 18.6%. From the definition of S given in Section 2, this means that the first screening reveals no suspicion of fraud for the other claims or that no loss is expected.

The other sample represents eighty percent of the claims, for which the

7

audit decision is left to the adjusters. The audit rate in this sample is close to five percent, which is much less than for the first sample of claims selected at random. The two samples have the same structure, which means that most of the suspicious claims are exempted from audit if the adjusters are left free to take the audit decision. The selection bias issue arises from the discrepancy between these two rates.

As indicated in the introduction, we will use the 51857 claims with no specific audit recommendation as the working sample, since they are subject to selection bias and are processed in a way currently used by insurance companies. The claims selected at random will be used as a holdout sample in order to assess the efficiency of the statistical model.

The fraud rate on audited claims is 19.8% in the working sample. The application to the hold-out sample of an estimated fraud equation estimated on the working sample does not modify the average fraud probability, which remains close to twenty percent. On the other hand, the similar fraud rate for the hold out sample is equal to 7.2%. Here we have an example of the overestimation induced by selection bias and mentioned in the introduction. A single equation model on fraud is unable to even partially fill the gap between the two fraud rates. Determining to what extent a selection model applied to the working sample can fill this gap is the purpose of the statistical analysis which follows.

4 The bivariate probit model: Theory and applications to selection bias assessment

4.1 The theoretical model

Using the notations from Section 2, we define the bivariate probit model on a sample of claims suspected of fraud. Referring to Figure 1, the bivariate model described below corresponds to the right part ($S = 1$) of the tree associated with a usual auditing strategy.

A bivariate model on the audit and fraud equation with a joint distribution on the two random components will be able to assess a fraud probability conditioned by the individual characteristics of the claims, but also by the audit variable A. Once this estimation is carried out, we have a fraud probability for suspicious claims which is unconditional with respect to an audit decision and which can be used in an optimal audit policy.

The bivariate probit model with censoring includes, first of all, an audit equation defined on all the claims which are suspected of fraud. This equation is defined in the following way

$$A_i = 1_{\left[A_i^* \geq 0\right]}; \; A_i^* = (x_A)_i \, \beta_A + (\varepsilon_A)_i \, ; \; (\varepsilon_A)_i \sim N(0,1) . \qquad (2)$$

The binary variable A is the sign indicator of a latent variable A^*. The variance of the random variable $(\varepsilon_A)_i$ can be set equal to one without loss of generality because of the invariance of A with respect to a multiplication of A^* by a positive constant. The regression components in the linear equation can be defined on the policy to which the claim is related or can be claim-specific. They are represented by a line vector whereas the parameters are stacked in a column vector which enables a cross product.

The fraud equation is defined on the audited claims only. We then write

$$\text{If } A_i = 1 : F_i = 1_{\left[F_i^* \geq 0\right]}; \; F_i^* = (x_F)_i \, \beta_F + (\varepsilon_F)_i . \qquad (3)$$

The random variable $(\varepsilon_F)_i$ also follows a standard normal distribution.

If we retain a bivariate normal distribution for $((\varepsilon_A)_i , (\varepsilon_F)_i)$, i.e.

$$\begin{pmatrix} (\varepsilon_A)_i \\ (\varepsilon_F)_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \; \rho \in [-1, 1], \qquad (4)$$

we obtain a bivariate probit model. The sign of the correlation coefficient ρ is of paramount importance for the estimation of fraud risk conditional on the audit variable.

In this censored setting, three levels are possible for the dependent variables. Let us compute the corresponding probabilities. We suppress the individual index for sake of simplicity.

1. If the claim is not audited, the fraud variable is not observed. We then have

$$P[A = 0] = P\left[\varepsilon_A < -x_A \beta_A\right] = \Phi(-x_A \beta_A) = 1 - \Phi(x_A \beta_A), \qquad (5)$$

where Φ is the distribution function of a standard normal variable and

$$\Phi(x_A \beta_A) = p_A = P[A = 1] = E(A).$$

The last equality in (5) results from the symmetry in the distribution of ε_A. The bivariate model is estimated on the suspicious claims only, hence the probability $P[A = 0]$ would be described as $P[A = 0 \,|\, S = 1]$ if expressed on all the incoming claims as we did in Section 2.

2. If the claim is audited and if fraud is established, we can write

$$P[A = 1, \ F = 1] = P\left[\varepsilon_A \geq -x_A \beta_A, \ \varepsilon_F \geq -x_F \beta_F\right]$$

$$= P\left[\varepsilon_A \leq x_A \beta_A, \ \varepsilon_F \leq x_F \beta_F\right] = P\left[\Phi(\varepsilon_A) \leq \Phi(x_A \beta_A), \ \Phi(\varepsilon_F) \leq \Phi(x_F \beta_F)\right].$$

We again used symmetry in the distribution of the random components. The variables $\Phi(\varepsilon_A)$ and $\Phi(\varepsilon_F)$ follow a uniform distribution on $[0, 1]$ and the distribution function of $(\Phi(\varepsilon_A), \Phi(\varepsilon_F))$ is a Gaussian copula indexed by ρ. If

$$p_F = P[F = 1] = P\left[\Phi(\varepsilon_F) \leq \Phi(x_F \beta_F)\right] = \Phi(x_F \beta_F),$$

we denote the Gaussian copula in the following way

$$C(\rho, p_A, p_F) = P\left[\Phi(\varepsilon_A) \leq p_A, \ \Phi(\varepsilon_F) \leq p_F\right] = P\left[\varepsilon_A \leq \Phi^{-1}(p_A), \ \varepsilon_F \leq \Phi^{-1}(p_F)\right].$$

This copula relates a bivariate Gaussian distribution function to its marginal components. Hence the probability of interest is equal to

$$P[A = 1, \ F = 1] = C(\rho, p_A, p_F). \tag{6}$$

3. The probability of the last level (claim audited but not fraudulent) is the complementary value

$$P[A = 1, \ F = 0] = P[A = 1] - P[A = 1, \ F = 1] = p_A - C(\rho, p_A, p_F).$$

Maximizing the log-likelihood on the sample requires a computation of the Gaussian copula and of its partial derivatives with respect to the parameters. This function does not have a closed form with respect to Φ and Φ^{-1} unless ρ is equal to the critical values $-1, 0$ and 1. It can be approximated by Gaussian quadratures (see Dionne, Gagné, Vanasse (1998) for an application to panel data with endogenous attrition).

We now present applications of the model to fraud probability prediction.

Under the null hypothesis $\rho = 0$, fraud probability does not depend on the audit variable and we have

$$p_F = P[F = 1] = P[F = 1 \mid A = 1]. \tag{7}$$

Under the alternative assumption $\rho \neq 0$, selection bias arises because in that case

$$P[F = 1 \mid A = 1] = \frac{C(\rho, p_A, p_F)}{p_A} \neq p_F. \tag{8}$$

10

Now these probabilities are confused in the application of a fraud risk model which does not take into account selection bias. Indeed, the conditional probability $P[F = 1 \,|\, A = 1]$ is estimated on the audited claims whereas the unconditional probability p_F should be applied on the incoming suspicious claims in order to design an audit policy. In order to have a better understanding of the influence of selection bias, basic properties of bivariate Gaussian copulas are recalled and applied in our context.

Three critical values for ρ are to be mentioned. First, we have

$$C(0, p_A, p_F) = p_A \times p_F$$

since the random variables ε_A and ε_F are independent if $\rho = 0$. In that case, all the fraud probabilities given in equations (7) and (8) are equal. Second, ε_A and ε_F are equal a.e. if $\rho = 1$, which implies $C(1, p_A, p_F) = \min(p_A, p_F)$. This value is the upper bound for a bivariate copula, the so-called upper Fréchet bound. Lastly, we have $\varepsilon_A = -\varepsilon_F$ a.e. if $\rho = -1$, hence $C(-1, p_A, p_F) = \max(p_A + p_F - 1, 0)$. Again, we reach the lower bound for a bivariate copula.

The map $\rho \longrightarrow C(\rho, p_A, p_F)$ is increasing on the interval $[-1, 1]$ for any value of (p_A, p_F)[1]. Hence the same result holds for $P[F = 1 \,|\, A = 1]$ in the bivariate probit model, whereas $P[F = 1 \,|\, A = 0]$ is a decreasing function of ρ.

4.2   Empirical applications of the bivariate probit model

The selection model should be applied to the only claims which are suspicious of fraud, which requires a variable pointing out these claims. Unfortunately the nature of the claims with respect to fraud suspicion is not available in our data base for the claims with no specific recommendation for audit. Among the 49290 claims which are not audited in the working sample (see Figure 2), some are suspicious of fraud ($S = 1$) and were exempted from audit due to a decision of the adjusters, and some are not suspicious. But the value of S is not available in the data base.

In order to apply the selection model described above, we will create a set of suspicious claims partly from simulation. This set must contain the 2567

---

[1] The distribution function of $(\varepsilon_A, \varepsilon_F)$ and hence the copula are integrals of the bivariate Gaussian density on negative orthants. Hence the increasing link with ρ is not surprising. We did not find a proof of this result in the statistical literature, but it is verified from numerical computations.

claims finally transferred to the SIU, because $A = 1$ implies $S = 1$. On the other hand, the proportion of suspicious claims should be the same on the working and the hold out sample since they have the same structure. We used an audit equation estimated on the hold-out sample which provides a suspicion probability conditioned on regression components since all the suspicious claims are supposed to be audited in this population. Then a random sampling scheme using this equation selected claims from the 49290 which are not audited. Together which the claims transferred to the SIU, they form a set of suspicious claims in the working sample which has the same weight as in the hold out sample (i.e. 18.6%). This approach is only a makeshift but we in any case find it interesting to estimate the selection model. Random auditing is not often carried out in the real world (which creates a selection bias problem), and a suspicion variable should be easy to determine during the first step of claims screening. Since we kept the proportion of suspicious claims observed with random auditing, the audit rate on suspicious claims in the working sample is $0.0495/0.186 = 0.266$.

Let us give an example to illustrate the importance of selection bias in relation with the bivariate probit model. We will compare the unconditional fraud probability of an incoming suspicious claim with average characteristics, and fraud probability if such a claim is audited with the audit policy used in the working sample. We expect different results from equation (8). We will compute $p_F = P[F = 1]$ as a function of ρ under the following constraints

$$p_A = P[A = 1] = 0.266; \ P[F = 1 \mid A = 1] = 0.198. \tag{9}$$

We just derived the proportion of audited claims among the suspicious ones, hence the value given to p_A. The frequency of fraudulent claims among those audited is 0.198, which explains the constraint on $P[F = 1 \mid A = 1]$. The unconditional fraud probability which we denote as $p_F(\rho)$ is a solution of the equation

$$C(\rho, p_A, p_F(\rho)) = P[A = 1, \ F = 1] = P[F = 1 \mid A = 1] \times P[A = 1]$$

$$\Leftrightarrow C(\rho, \ 0.266, \ p_F(\rho)) = 0.198 \times 0.266 \simeq 0.0527. \tag{10}$$

Since the claims adjusters are supposed to be able to capture idiosyncrasies in fraud distributions which are not summarized by the observable information, we expect that

$$p_F(\rho) = P[F = 1] < P[F = 1 \mid A = 1] = 0.198.$$

12

Indeed, the unconditional probability of fraud $p_F(\rho)$ is a weighted average of the two conditional probabilities $P[F = 1 \mid A = 1]$ and $P[F = 1 \mid A = 0]$. The latter probability should be lower than the first one because of the information brought by the audit decision. In that case, ρ should be positive since we have

$$\rho > 0 \iff p_F(\rho) < 0.198$$

from the equivalence

$$\rho > 0 \iff 0.198 \times 0.186 = C(\rho,\, 0.186,\, p_F(\rho)) > C(0,\, 0.186,\, p_F(\rho)) = 0.186 \times p_F(\rho).$$

We used the increasing property of the copula with respect to ρ.

Let us give values for $p_F(\rho)$ under the constraints given in (9) and (10). We let ρ increase from 0 to 1 with an increment equal to 0.1. Results are presented in Table 1.

Table 1: Unconditional fraud probability for an incoming suspicious claim with average characteristics, as a function of the correlation coefficient

ρ	$p_F(\rho)$
0	0.198
0.1	0.166
0.2	0.140
0.3	0.117
0.4	0.099
0.5	0.084
0.6	0.072
0.7	0.063
0.8	0.056
0.9	0.053
1	0.053

Suppose that $\rho = 0.5$. The fraud probability of an incoming suspicious claim with average characteristics is 2.4 times less than if this claim is audited. This would drastically modify the threshold related to a score designed to select claims for audit.

From Table 1, the unconditional fraud probability of an incoming suspicious claim with average characteristics is equal to 7.2% (the fraud rate observed

without selection bias in our data base) for a value of ρ close to 0.6. Hence a positive estimated correlation coefficient is expected on the working sample.

This estimated coefficient depends on the choice of regression components in the bivariate probit model. On the whole, we noticed that $\widehat{\rho}$ decreased with the amount of information used in the regressions. Variations of $\widehat{\rho}$ due to the sampling scheme of suspicious claims in the working sample are much less important than those due to the choice of regression components in the bivariate model.

Let us detail for instance estimation results with a medium number of regression components. All of them are significant at a one-percent level. With this constraint, the set of covariates retained in each equation cannot be increased with the variables at our disposal. Maximum likelihood estimations on the equation generating the sampling scheme and in the bivariate probit model are given in Table 2.

Table 2: Regression results in the bivariate probit model (working sample) and in the equation on fraud suspicion (hold out sample)

	Sampling scheme	Bivariate probit model	
Equations	Suspicion	Audit	Fraud
Sample	Hold out sample	Working sample, $S = 1$	Working sample, $A = 1$
Size of the sample	12928	$\simeq 9640$	2567
Parameters	$\widehat{\beta_S},\ (S = A)$	$\widehat{\beta_A}$	$\widehat{\beta_F}$
Intercept	-1.53	-1.08	-1.90
Seniority of policyholder:			
less than one year			0.21
Vehicle=motorbike			0.41
Automobile, private use			0.39
Number of previous claims			0.06
Coverage: Third party liability only	0.71	0.74	
Coverage: Third party liability + theft, arson and glasses	0.66	0.53	
Third party at fault	0.33		
Use of the no-fault system	0.29		
Age of the policyholder	-0.003	-0.005	

14

The results depend on the regression components but also on random sampling for the bivariate probit model. The estimators are stable with respect to suspicious claims sampling. With this set of regression components, we have

$$\widehat{\rho} \simeq 0.51.$$

Besides, the estimated unconditional fraud probability for incoming suspicious claims is 8.4% on average. With this set of regression components, the selection model provides a satisfactory result, since the goal is to reach a 7.2% ratio from a starting point - the fraud rate for audited claims in the working sample - which is 19.8%.

As mentioned above, the estimated correlation coefficient increases if fewer regression components are retained. The estimated coefficient ranges between 0.36 and 0.64, depending on the number of regression components. The average unconditional fraud probabilities range between 6.9% and 10.8%.

The positive result obtained in this section is that the selection model is able to get very close to the actual fraud rate, which can only be reached through random auditing. The weakness of selection models applied to censored data is that estimation results highly depend on the regression components set.

5 Applications of selection bias models to auditing policies design

An audit decision based on a short term analysis compares estimated audit cost and gain from fraud detection. The gain is the product of the claim settlement reserved cost and of the fraud probability[2] (see Ayuso, Guillén et al. (2004) for the inclusion of costs in audit policies). The reserved cost of claim settlement is determined by the adjuster during the first screening of the claim. Hence it is known by the insurance company before the audit decision, which is not the case with the audit cost. The cost recorded in the data base and related to audit corresponds to all the steps of the claim examination, including the first screening.

Let us denote the audit cost related to the SIU examination for claim i as ac_i, and c_i the reserved claim cost. A transfer of this claim to the SIU

---

[2]If the claim is proved fraudulent, we suppose that the agreement with the policyholder consists in cancelling the compensation. We discard other possible long-run consequences, such as non-renewal of the policy.

generates an expected gain in the short run if

$$\widehat{E}\left[c_i F_i - AC_i\right] > 0 \Leftrightarrow \widehat{E}[F_i] = \widehat{P}[F_i = 1] > \frac{\widehat{E}[AC_i]}{c_i}. \qquad (11)$$

Indeed, $c_i \widehat{E}[F_i]$ is the expected gain from audit if claim i is transferred to the SIU. The expected fraud probability is derived from a bivariate probit model in what follows. The audit cost is only known *ex post* (hence its random variable status at this step of the computation).

Information on audit costs is necessary in order to derive the expected values $\widehat{E}[AC_i]$. The average audit costs are given in Table 3 depending on the status of the claim and on the sample.

Table 3: Average audit costs (including the first screening)

Average audit costs of claims	Hold-out sample $(A = S)$	Working sample $(A \neq S)$
$A = 0$	€ 36.97	€ 37.73
$A = 1$, $F = 0$	€ 59.81	€ 67.82
$A = 1$, $F = 1$	€ 231.71	€ 222.84

These results clearly indicate that the more a claim is likely to be fraudulent, the more it must be examined thoroughly by the adjusters, which increases the audit cost. The averages are similar in both samples. The difference between the averages of audit costs for claims transferred to the SIU and not fraudulent can be explained by the more severe selection for SIU in the working sample, which makes these claims more suspicious of fraud. Hence selection bias also appears for audit costs.

The audit costs in Table 3 correspond to all the steps of claims examination, including the mandatory first screening which is not performed by the SIU. The average audit cost charged to the SIU will be set equal to

$$\overline{ac}^{NF} = 67.82 - 37.73 = € 30.09; \; \overline{ac}^{F} = 222.84 - 37.73 = € 185.11, \qquad (12)$$

where $\overline{ac}^{NF}$ and $\overline{ac}^{F}$ relate to non fraudulent and fraudulent claims. This computation implicitly supposes that the cost of the first screening does not depend on the eventual status of the claim.

The expected audit cost clearly depends on fraud probability since $\overline{ac}^F > \overline{ac}^{NF}$. Conditioning this expectation on average values for each level of the audited claims, we obtain

$$\widehat{E}[AC_i] = \left(\widehat{P}[F_i = 1] \times \overline{ac}^F\right) + \left(\widehat{P}[F_i = 0] \times \overline{ac}^{NF}\right)$$

$$= \overline{ac}^{NF} + \left(\widehat{P}[F_i = 1] \times (\overline{ac}^F - \overline{ac}^{NF})\right),$$

An audit policy can be designed from this assessment of the audit cost. The rule given in (11) suggests transfer claim i to the SIU if

$$c_i > \overline{ac}^F - \overline{ac}^{NF} \ \& \ \widehat{P}[F_i = 1] > \frac{\overline{ac}^{NF}}{c_i - (\overline{ac}^F - \overline{ac}^{NF})}. \tag{13}$$

Since the probability threshold must be less than one, the first condition amounts to $c_i > \overline{ac}^F$. We will use the hold-out sample to compare the audit policies derived from different fraud models. All the suspicious claims are audited if random auditing is performed thoroughly. In that case there is no selection bias and the target audit rate is obtained from a fraud equation estimated on the audited claims in the hold-out sample. From the fraud equation specified in the bivariate model of Table 2, and with the selection rule and the audit costs given in (13) and (12), we obtain an audit rate for suspicious claims of 36.7%.

Let us assess the influence of selection bias on audit policies. Suppose that the fraud equation specified in Table 2 is estimated on the audited claims of the working sample. The optimal audit rate on suspicious claims of the working sample goes up to 62%, which reflects the overestimation of fraud probability if selection bias is neglected.

Let us see how the target audit rate obtained in the first place can be approximated by a selection model. The proportion of suspicious claims which should be audited in the working sample is 39% if we use the selection model estimated in Table 2. This audit rate is rather close to the target rate obtained from random auditing (i.e. 36.7%), which is a satisfactory result.
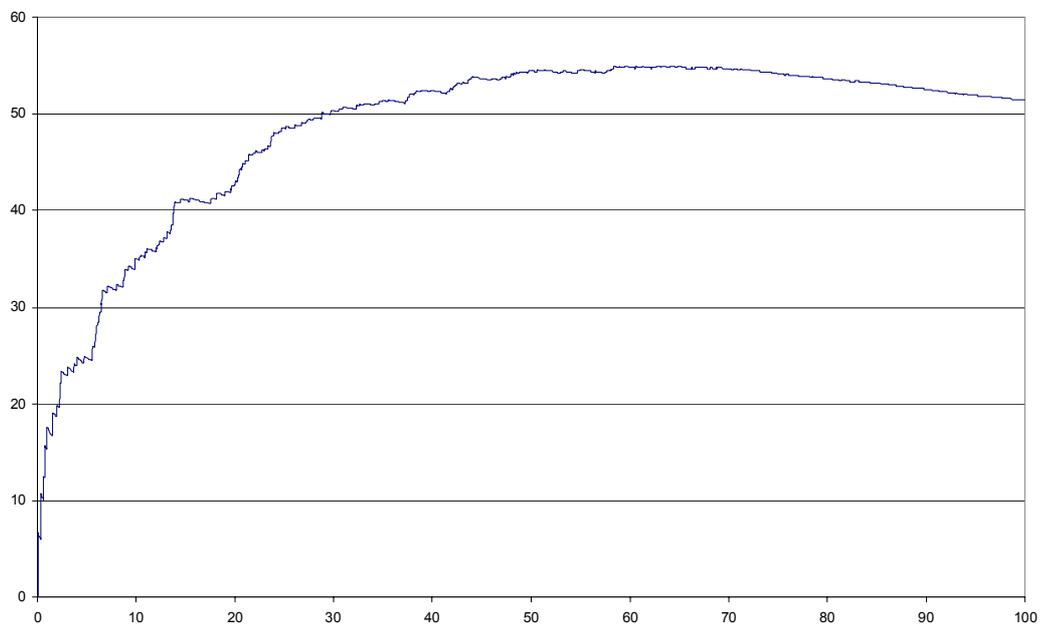
If the set of regression components varies in the regression, the optimal audit rate ranges between 33 and 45% in our different trials. These rates strongly depend on the set of regression components through the estimated correlation coefficient. This result is obviously negative since sound risk assessment derived from random auditing cannot be precisely anticipated from the selection

model. However the bivariate probit model points out selection bias and can justify a random auditing strategy which is not often employed by insurance companies.

The efficiency of these audit policies is almost the same for all of them, which may seem surprising since the audit rates exhibit great variations. As for the selection rule given in (13), we compare these policies to a zero audit strategy. Hence gains are the reserved costs of the audited claims proved fraudulent in the hold-out sample, and losses are the audit costs charged to the SIU. For all the preceding audit policies, the average gain per suspicious claim is close to $\in$ 52. This stability with respect to audit rates is explained by the shape of efficiency curves. If claims are sorted in decreasing order with respect to the expected gain $\widehat{E}[c_i F_i - AC_i]$, the average gain per suspicious claim expressed as a function of the audit rate provides such a curve. Figure 3 is related to the bivariate model estimated in Table 2.

All the efficiency curves have the same shape. This one reaches its maximum for an audit rate close to 60%. The stability of the curve for audit rates greater than 30% explains why the different policies have similar efficiencies.

Figure 3: Average gain (euros) per suspicious claim as a function of the percentage of audited claims

6   Conclusions

Let us first comment briefly on the instability of selection bias assessment with respect to the set of regression components. This instability is actually consubstantial with the censored character of the data. We presented censored and non-censored selection models in the introduction. Such models are made up of a selection variable (e.g. audit), a variable of interest (e.g. fraud) and are censored if the variable of interest is observed only on the selected individuals. For a non-censored model, the influence of the selection variable on the variable of interest is easily derived from the comparison of two samples (selected versus not selected individuals) with respect to the variable of interest. In a censored context, this influence is only assessed through the variation of estimated selection probabilities on the selected individuals. This estimated probability plays the role of a supplementary covariate in the regression model related to the variable of interest. Now this probability is derived from the information already used in the regression. Selection bias reflects a more intricate specification for the distribution related to the variable of interest instead of being based on observed differences in the non-censored setting[3]. If the selection model is unstable for censored data, it is however of greatest interest precisely in this context.

Since random auditing cannot be avoided in order to obtain a sound assessment of fraud probabilities, we find it interesting to derive the cost of this controlled experiment on our data base. The relative costs of the audit strategies in the two samples described in Figure 2 can be derived with respect to a situation where no claim is transferred to the SIU, which we did in the last section. Hence gains are the reserved costs of the audited claims proved fraudulent, and losses are the audit costs charged to the SIU.

The result is striking. The controlled experiment (i.e. random auditing) actually generates more gain for the insurance company than the usual audit strategy, regardless of the possible future gains induced by better fraud probability assessment. Hence the insurance company would probably have saved money had all the suspicious claims of the working sample been transferred to the SIU. In addition, increasing the audit rate reduces fraud risk because of the deterrence effect of audit policies (see Tennyson and Salsas (2002); Dionne,

---

[3]The seminal paper on selection models with censoring is entitled "sample selection bias as a specification error" (Heckman, (1979)). Estimated selection probability is included in a linear regression on the variable of interest via the inverse Mills' ratio, which creates the so-called "Heckit".

Giuliano and Picard (2002)).

Lastly, let us mention other issues which must be considered in the design of an optimal audit policy. First, the anticipation of audit costs could easily be improved since our analysis remained at a basic level. More important is the impact of claims settlement decisions on the policyholder's value. The thorough audit of a claim may modify the relationship between the policyholder and the insurance company. If the claim is not found to be fraudulent, the insured may take the audit process amiss and the attrition rate should rise as a consequence. If the claim is found fraudulent, the risk level of the policyholder will be updated at a higher level. A fraud event usually triggers an increase in premium or a cancellation of the policy. Taking into account the policyholder's value in an audit decision would increase both gains and losses computed in the short run in Section 5. This would also imply a long-run econometric analysis from the policies data base.

References

[1] Artis, M., M. Ayuso and M. Guillén, 2002, Detection of Automobile Insurance Fraud with Discrete Choice Models and Missclassified Claims, *Journal of Risk and Insurance*, 69(3): 325-340.

[2] Ayuso, M., M. Guillén, S. Viaene, and D. Van Ghee, 2004, Cost-Sensitive Design of Claim Fraud Screens, *Lecture Notes in Artificial Intelligence*, 3275, 78-87.

[3] Belhadji, E.B., G. Dionne and F. Tarkani, 2000, A Model for the Detection of Fraud, *Geneva Papers on Risk and Insurance - Issues and Practice*, 25(5): 517-538.

[4] Chiappori, P.A. and B. Salanié, 2000, Testing for Asymmetric Information in Insurance Markets, *Journal of Political Economy*, 108(1): 56-78.

[5] Dionne, G., R. Gagné and C. Vanasse, 1998, Inferring Technological Parameters from Incomplete Panel Data, *Journal of Econometrics*, 87: 303-327.

[6] Dionne, G., F. Giuliano and P. Picard, 2002, Optimal auditing for insurance fraud, Working paper, available at http://www.hec.ca/gestiondesrisques/cahiers.htm .

[7] Hand, D.J. and W.E. Henley, 1997, Statistical Classification Methods in Consumer Credit Scoring: A Review, *Journal of the Royal Statistical Society A*, 160: 523-541.

[8] Hausman, J.A. and D.A. Wise, 1979, Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment, *Econometrica,* 47(2): 455-474.

[9] Heckman, J.J., 1979, Sample Selection Bias as a Specification Error, *Econometrica,* 47(1): 153-162.

[10] Heckman, J.J., 2000, *Microdata, Heterogeneity and the Evaluation of Public Policy*, Nobel Prize Lecture, available on the Web site `http://nobelprize.org/`.

[11] Rosenbaum, P.R. and D.B. Rubin, 1983, The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70: 41-55.

[12] Tennyson, S. and P. Salsas-Forn, 2002, Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives, *Journal of Risk and Insurance*, 69(3): 289-308.